ENHANCING CLOUD PERFORMANCE WITH CLARA AND NATURE-INSPIRED PARADIGM

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies
(Deemed to be University)

by

TANVI GUPTA
14/Phd/004

Under the supervision of
Dr. SUPRIYA P PANDA
Professor, Faculty of Engineering &Technology

FACULTY OF ENGINEERING & TECHNOLOGY

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

SEPTEMBER 2018

FACULTY OF ENGINEERING & TECHNOLOGY

Certificate

I hereby declare that the short synopsis "**ENHANCING CLOUD PERFORMANCE WITH CLARA AND NATURE-INSPIRED PARADIGM**" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in COMPUTER SCIENCE & ENGINEERING under faculty of Engineering & Technology of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2014-15.

It is a bona fide record of original work carried out under the supervision of **Prof.(Dr.)SUPRIYA P PANDA** and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

TANVI GUPTA
(Name of the Scholar)
14/Phd/004

Signature of the Supervisor
DR. SUPRIYA P PANDA
PROFESSOR
FACULTY OF ENGINEERING & TECHNOLOGY

(Chairman, DRC)

Chairperson, CDP

ABSTRACT

Cloud Computing is the structured model in IT, in which the services in the form of infrastructure, platform or software are pooled together and are provided to the users by Cloud providers on the pay-as-needed basis. The sharing of resources may initiate a problem of availability, as the demand of these services or resources are increasing day by day, it is being difficult to handle the load by the cloud providers at such a high rate. Here, the load balancing concept comes into existence. Load balancing means to distribute the workload among all virtual machines (VMs) in a manner that all of them are fully utilized but not over utilized. If any of the virtual machine (VM) is over utilized, the task will be removed from that virtual machine and given to an underutilized virtual machine. It includes the two concepts: task scheduling and high resource utilization. Here, task scheduling and management of the resources are closely related to each other. For this whole scenario, Nature-inspired Meta-heuristic Algorithm and Multi-Objective Task Scheduling with clustering concepts are used. As it is known that Load Balancing is a NP- Hard problem, so it can be solved easily by Meta-heuristic Nature-inspired techniques. To further improve the performance of cloud environment, clustering of the cloudlets and virtual machines is to be done. Moreover, instead of just considering a single Quality of Service (QoS) parameter for clustering, Multi-QoS parameters of each virtual machines and cloudlets are being considered so that better results may come. So, the performance of the cloud environment depends on the balancing of load, which includes certain factors like response time, makespan and cost. So, the aim is to enhance the performance of the cloud environment by reducing the response time, makespan and cost. This will help to achieve higher user satisfaction and resource utilization.

Key words: Load balancing, Clustering, Nature-inspired, virtual machines, Meta-heuristic, NP-Hard Problem, Multi-Objective.

CONTENTS

S. No.	Description	Page No.
1	**Introduction** **1.1 Cloud Computing** **1.2 Characteristics of cloud computing** **1.3 Architecture components** **1.4 Cloud deployment model** **1.5 Advantages of Cloud Computing** **1.6 Disadvantages of cloud computing** **1.7 Issues and Challenges in cloud computing**	**1-9**
2	**Literature Review**	**10-15**
3	**Description of Broad Area/Topic**	**16-23**
4.	**Problem Identification/Objective of the Study**	**24**
5	**Methodology to be adopted**	**25**
6	**Proposed/expected outcome of the research**	**26-27**
7	**Proposed Time Frame (Gantt Chart)**	**28**
8	**References**	**29-30**

1. Introduction

1.1 Cloud Computing

"Cloud Computing" is today's technology, which provides the technical resources and information or software that are required by the end user. The user will pay to the cloud provider according to the use of the resources. Resources may be infrastructure, platform or software. In fact, in cloud computing there is a shift in technology based on social, ethical, economical and legal implications. Here, the end users avail the services, which are provided by the providers through datacenter and moreover store the data privately on the systems that they do not have. Figure 1 depicts Overview of Cloud Computing.

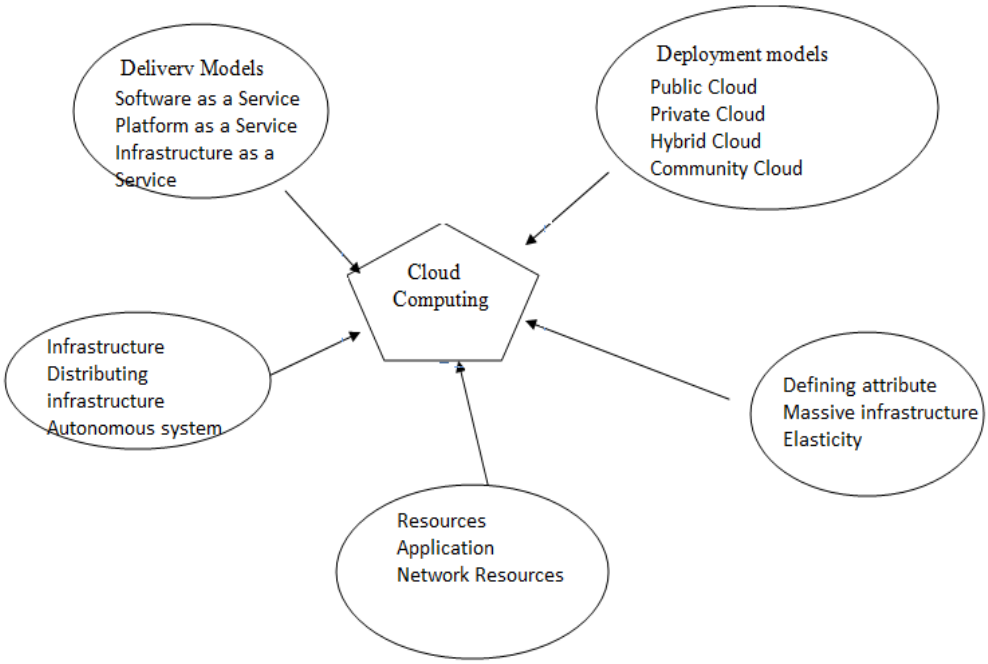

Figure 1: Overview of Cloud computing

There are three principle aspects for a Cloud service differs from traditional hosting (Khazaei et al., 2012). Firstly, On demand it is provided ; secondly, elasticity as the users can use the services as much as or as little as they want at any given time(typically by minute or the hour); and thirdly, the provider(Amit et al., 2013) fully managed the services. The assumption is that any job or task when received by the cloud center is to be serviced by the appropriate node selected by the load balancer so that there is balance in the network; after the job is finished the job or task will leave the node. Also, Table 1 shows the Challenges and Scope in Cloud Computing Paradigm.

Table1: Challenges and Scope in Cloud Computing Paradigm

	Challenge	Scope
1	Availability of service	Use Multiple cloud vendors, Use Elasticity to prevent Distributed Denial of Service
2	Data Lock-In	Standardize APIs; Compatible SW to enable Surge Computing
3	Data Confidentiality and Auditability	Encryption deployment, VLANs, Firewalls, Geographical Data Storage
4	Bottlenecks in Data Transfer	FedExing Disks; Data Backup/Archival; Higher BW Switches
5	Unpredictability in Performance	Improved VM Support; Flash memory;
6	Storage Scalability	Invent Scalable Store
7	Bugs in Large Distributed Systems	Invent Debugger that relies on Distributed VMs
8	Quickly Scaling	Invent Auto –Scalar that relies on Machine Learning; Snapshots for conversation
9	Fate Sharing Reputation	Offer reputation –guarding services like those for email
10	Software Licensing	Pay –for-use licenses; Bulk use sales

1.2 **Cloud Computing Characteristics**

a) On demand self services: Computer services such as email, applications, network or server service can be provided without requiring human interaction with each service provider. Cloud service providers providing on demand self services include Amazon Web Services (AWS), Microsoft, Google, IBM and Salesforce.com. New York Times and NASDAQ are examples of companies using AWS (NIST). Gartner describes this characteristic as service based.

b) Broad network access: Cloud Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms such as mobile phones, laptops and PDAs.

c) Resource pooling: The provider's computing resources are pooled together to serve multiple consumers using multiple-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The resources include among others storage, processing, memory, network bandwidth, virtual machines and email services. The pooling together of the resource builds economies of scale.

d) Rapid elasticity: Cloud services can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the

capabilities available for provisioning often appear to be unlimited and can be purchased in any measure at any point in time.

e) Measured service: Cloud computing resource usage can be calculated controlled, and the details would provide transparency of the utilized service for both the provider and the user. To control and optimize the use of resources, metering capability is used by cloud computing services. This means that as air "time", "electricity" or "municipality water", IT services are charged per usage metrics.

f) Pay per use: Bill is according to our utilization of resources, the more we utilize, and the higher the bill is. Now a days, just as the telephone services sell the voice and data services, IT providers sell data center hosting, network security management etc, so we can say it is easy to measure the quantity of the services used and pay according to that bill.

g) Multi tenacity: "Multi tenacity" is a crucial characteristic of "cloud computing", which is controlled by "Cloud Security Alliance". For different consumer constituencies, it serves the requirement for "policy-driven enforcement", "segmentation", "isolation", "governance", "service levels", and "chargeback/billing models". Consumers share infrastructure, which may use a "public cloud" provider's service contributions or in reality be from the identical group.

1.3 Architecture Components

The "Cloud service models" are as follows: "SaaS","PaaS", and "IaaS". Figure 2 represents "Cloud Service Models" (B. Grobauer et al., 2011). For the analysis of "Security" issue, the architecture shown below is endowed with a conceptual summary of "Cloud Computing" that makes maximum "Security-Relevant" cloud resources explicit.

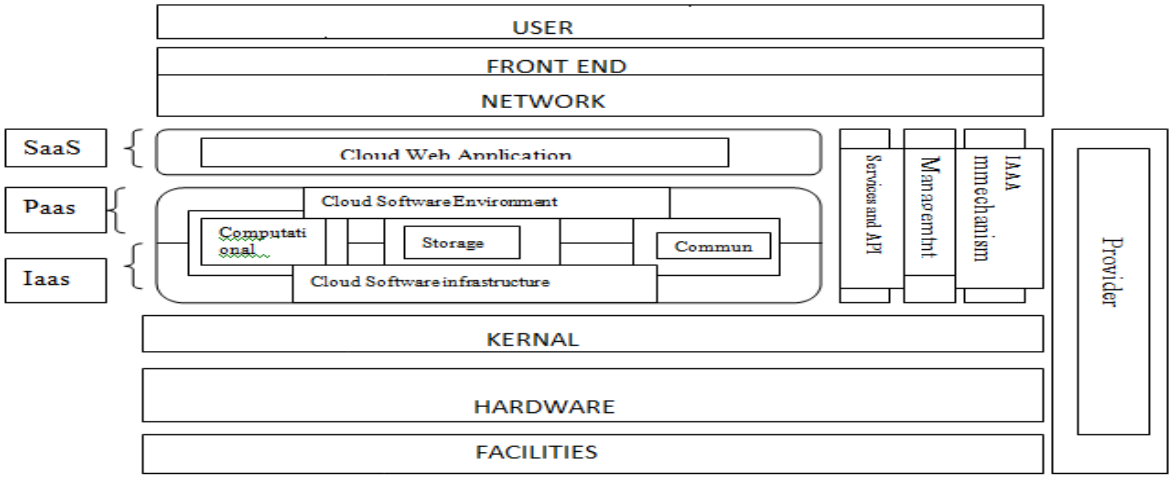

Figure 2: Cloud Service Model

A. Software as a Service (SaaS)

"SaaS" (B.Grobauer et al., 2011) refers to as "Software as a Service" which be a nomenclature of "Cloud Computing". This service includes the "software" as a whole provided to the user by the provider. The users are charged according to the use of the software not for the platform and infrastructure given to the user to support the software. Examples can be DBMS software, MIS, Payroll processing software, Tally Software etc.

B. Platform as a Service (PaaS)

To build up "Cloud Services" and "Applications" (e.g. "SaaS") directly on the "PaaS" cloud by cloud consumers. "PaaS" (B.Grobauer et al., 2011) is one of the other categories of cloud service which provides the user a platform with no difficulty of structuring and upholding the infrastructure. The users are charged according to the use of platform in particular time they are not charged for the infrastructure used in it. Example of "PaaS" is "Google AppEngine".

C. Infrastructure as a Service (IaaS)

"IaaS" (B.Grobauer et al., 2011) is the cloud computing service, which provides the infrastructure to the users. These services are provided by the "cloud provider" to the "cloud users" according to their need and demand. The cloud users are then charged according to the infrastructure they used in particular time. The infrastructure may involve hardware, storage, network and processor. Example of "IaaS" is "Amazon's EC2".

1.4 Cloud Computing Deployment Model

1.4.1. Public Cloud

A "Public Cloud" is the cloud service model for the general public that can be used by them for various purposes, which needs no privacy. This cloud is owned by "third party cloud provider". Mostly this type of cloud is used as general and is based on pay –per –usage model. The general population uses this cloud irrespective of any organization affiliation. Figure 3 shows that the cloud users, which is actually an organization, accesses the services provided by the cloud provider.

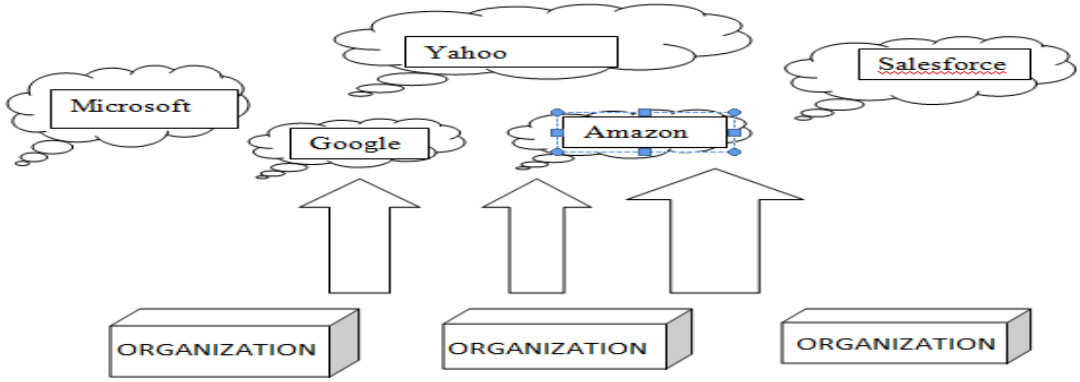

Figure 3: Overview of Public Cloud

1.4.2. Private Cloud

A "Private Cloud" is the cloud used by particular organization so that privacy can be maintained from outside sources. So, for this organization buys own hardware and software to set a private cloud. A Private Cloud thus provides the security and privacy of data so that the data will remain in the organization network. For the Private Cloud there should be some trust boundaries for the organization. In the Private Cloud, an administration can be passed out by internal or outsourced staff. Figure 4 shows the organization which the cloud service consumers access the cloud privately using virtual private network.

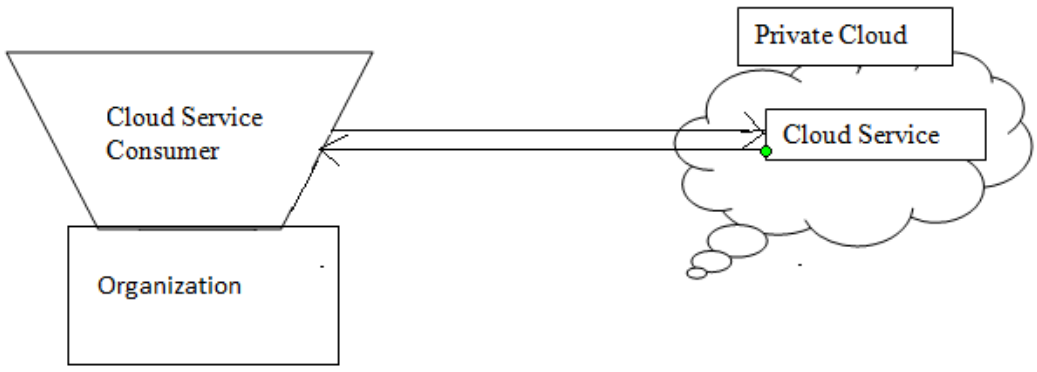

Figure 4: Overview of Private Cloud

1.4.3. Hybrid Cloud

A "Hybrid Cloud" is one of the "deployment models", which includes the advantages of both public and private cloud. As there is only some data that need to be confidential, for which private network is needed and otherwise public cloud is fine. For this an example can be cited; as

a Bank uses both public and private cloud; mainly the transactions are done on private cloud and rest of the work is done in the public cloud, depicted in Figure 5.

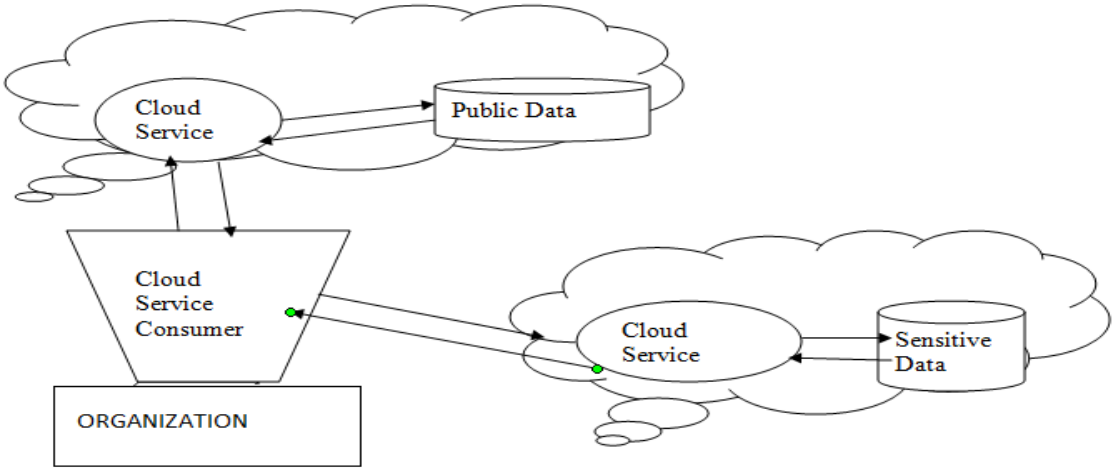

Figure 5: A 'Hybrid Cloud Architecture' having both public and private cloud.

1.4.4. Community Cloud

"Community Cloud" is employed by a specific community of people, which can access the public cloud but with limited access. A community cloud is mainly for the specific people they create a group for which they need. Also, if a person of community is a member of community that does not mean that he has all the rights of the group to access any IT resources. One more important thing is that, unless allowed by the community, parties outside the community are not granted access. To define and to evolve the community cloud member, the cloud consumers of the community usually divide the job among themselves. Figure 6 shows the concept of Community Cloud.

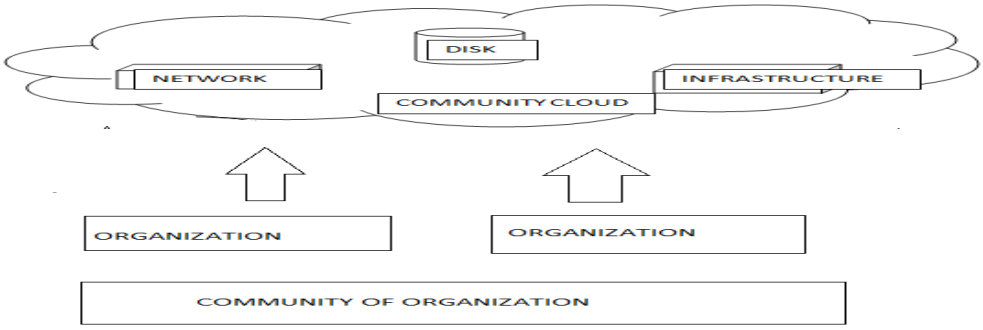

Figure 6: Overview of Community Cloud

1.5 Advantages of Cloud Computing

1) **Cost efficiency:** – The main benefit to be considered is less cost than an on-premise technology. As cloud offers enormous space or storage so, so the organizations or companies do not use the disks to store the data which helps in saving money and resources.
2) **High speed:** – High speed means that the resources are available to the users within a minute through Cloud computing. We can deploy the services at a very fast pace through this.
3) **Excellent accessibility:** – When you store the information in the cloud it means it can be accessible anywhere, anytime regardless of the location of the machine.
4) **Back-up and restore data:**– Backup and restore data can be accessed easier if the data is once store in cloud otherwise, it is a time taking process

1.6 Disadvantages of Cloud Computing

- **Security issues:** – The "Security" of the data is a main concern when the data is once on the internet. Though there are various algorithms for security but still thefts of the data can be done.
- **Low bandwidth:** –Cloud computing can be done only if the speed is good but if there is lower bandwidth then automatically with many users, the speed will degrade. So, low bandwidth is one of the disadvantages of Cloud Computing.
- **Flexibility issue**s: – As in cloud computing the services are to be provided to the user by the cloud provided so these services have to run on the remote system due to which it may not work properly so, flexibility is one of the alarm in "Cloud Computing".
- **Incompatibilit**y: –As the services that are given by the provider are in the virtualized format, so incompatibility can be possible.

1.7 Challenges & Issues in Cloud Computing

A. Security and Privacy

The three concerns in adoption of cloud are security, performance, and availability .As the data is present in cloud it can be accessed or modified by the cloud providers illegally. Also, there are two types of security threats that are external and internal. In external risk can be posed by the

outside people of the organization, like hackers etc. and the internal threats can be posed by the third party people that have an access over the network of the organization also the contractor, former employees etc. So, there are some “Security and Privacy” concerns in Cloud Computing.

B. Performance

“Performance” is the prime challenge in “Cloud Computing”. Most of the research in Cloud Computing is based on performance. If the performance degrades it will lead to the loss of customers. Also, end of service delivery also reduces the bottle line revenue. Reasons of “reduced performance” are due to dearth of resources like “disk space”, “limited bandwidth”, “lower CPU speed”, “memory” etc.

C. Reliability and Availability

“Reliability and Availability” are the strengths of technology which means that the resources should be available anywhere, everywhere and at anytime without any loss of data. Reliability is one of the crucial features of cloud computing but can be hampered by downtime. If the particular server is down at the time of access by some user the data cannot be retrieved by the user. To achieve reliability in these resources can be duplicated and can be used at the downtime. Another aspect is availability i.e. data can be accessed anywhere at any time. But due to low bandwidth or no internet availability etc. resources or data can’t be reached to the user during access.

D. Scalability and Elasticity

Scalability and Elasticity is the topic of concern in cloud computing. Scalability means to scale up the resources according to the need in the cloud environment and elasticity means to scale up and down according to the need. Here, elasticity states the scalability. Also, scalability is of two types one is horizontal and one is vertical. Horizontal scalability means new system will be added into the environment whereas vertical means resources are added in the environment.

E. Interoperability and Portability

Interoperability means exchange of information through computer systems. Interoperability has become a challenge in cloud computing due to vendor lock-in risk. There are various level of

"Interoperability" that are application, service, management and Data interoperability. The major concerns are the lack of open standards, open APIs and lack of standard interfaces for VM formats and service deployment interfaces. Cloud Portability identifies one cloud solution, which will be able to work with other platforms and applications along with other clouds.

F. Resource Management and Scheduling

Quality of Service (QoS) is a vital factor of the cloud computing. Hence, a quality should be upheld by the services provided by the cloud provider to the cloud user. So, for quality resources should be managed and the tasks should be scheduled. Resource management includes the management of the resources like hardware, storage, network and processor, these should be managed in such a way by the providers through virtualization that every user will get the quality service. And in job scheduling following parameters are to be considered that are priority, whether the jobs are preemptive or non preemptive etc. The main thing is to get the minimum execution time, minimum makespan and maximize throughput.

G. Energy Consumption

On example here to be considered is the survey done by Amazon. It states that the approx. 50% of the cost is consumed by the server for the management of three years while the power consumption i.e. energy and cooling requirements consume approx 40%.

2. Literature Review

"Load Balancing" is one of the concepts in cloud environment to balance the workload in the environment. Here the process goes on without any system imbalance and the resources are fully utilized. For this, there are various proposed load balancing algorithms. Many of the researchers are doing the work on balancing load to enhance the performance of the cloud environment. These are mentioned as below:

Mohammad Sajid et al. (2013) anticipated the "Round Robin load balancing algorithm", Authors discussed that while processing the data a "time slice" is maintained. The mechanism is as follows: In this, every process would be executed within the quantum slice and then switch to another process. All this would be in a ring manner. A balance technique is followed until all the processes complete their task, according to the particular sequence to "balance" the method in a group. One of the good examples is the web servers which uses http request to distribute equally as these requests are in similar manner. The disadvantage of this algorithm is that the algorithm gives good result only if all the servers have the same specification.

Poonam Devi et al. (2013) proposed the "Shortest Job First scheduling" (SJF) algorithm .In this, the approach is to utilize the resources fully for heavy jobs by performing the complete execution of short jobs so, in this shortest executable job is first selected. As the processes have to wait less, so the execution of the job which is the shortest will enhance the on the whole performance of the system. The main consequence of this algorithm is starvation which means a task may have to wait for very long time for some requested resource.

S.C. Wang et al. (2010) anticipated an algorithm combining OLB+LBMM. Its main aim is to attain better execution efficiency of the machine. "OLB algorithm" here puts each single machine in working mode so that no node remains idle and the goal of cloud computing is achieved. But the disadvantage of this algorithm is that it does not consider execution time of node. Similarly, LBMM algorithm will help in plummeting "execution time" of the jobs on machine, that lessen the on the whole "completion time". So, hybridizing both algorithms will result in better performance of cloud computing. .

Bhadani et al. (2010) anticipated the policy named CLBVM which states that the load should be distributed to the "Virtual Machines" in a way that the environment remains balanced. This policy will enhance the fault tolerance as everything is dependent on central system.

J. M. Galloway et al. (2011) anticipated the **"**Power Aware Load Balancing Algorithm"**.** In this algorithm, "utilization percentage" is calculated for the modules which are in the working condition and according to that the operating computing nodes are decided and others that do not work will get shut down. In working module there are three sections: "Balance section", "Upscale section", and "Downscale section". "Balance section" is the section where the VM is going to start for determining initialization process, the second section is the" upscale section" in which additional computing nodes will power-up and the third section is" downscale" in which the idle node will get shut-downs in the process. On the comparison with the other existing algorithm based on the consumption this algorithm works.

Srinivas Seth et al. (2012) anticipated the "Fuzzy Active Monitoring Load Balancing (FAMLB)"**.** In this, there are two parameters that are considered of the virtual machine in this algorithm 1) processor speed and 2) load. Along with the concept of fuzzy logic and the "dynamic load balancing algorithm" there are certain extra parameters need to be considered that are "memory usage", "bandwidth usage", "disk space", and "virtual machine status".

S.Wang et al. (2010) In the paper, the authors state that in "throttled load balancing algorithm" the "load balancer" creates an "index table" for the "virtual machine". In this table the state of the "virtual machine" is saved i.e., available or busy. Now, to perform the job, the client/ server will ask the data center for the idle virtual machine for the task. Then, the load balancer will be queried by the data center for the allocation of "Virtual Machine". After that the "Load Balancer" will find the first available virtual machine in the index table. If VM is found, the data center is loaded. But, this algorithm works only when hardware configurations of all VMs of data centers have similar hardware configurations.

M. Randles et al. (2010) suggested the active clustering algorithm, which says that similar nodes are grouped together and work on the same principle. In this the performance is improved by increasing the parameter outcome and degraded thereby if there are heterogeneous parameters or there is no diversity in the system. In this process, a matchmaker node is selected from the neighbor, which will satisfy the criteria that the former node should be different from neighbor node. The process execution is as follows:

i. In the First step the connection will be formed between both matchmaker and neighbor node, one of the constraints in it is both should be of similar to the initial node.

ii. In the Second step the "matchmaker node" will be detached from the connection "neighbor node".

M. Randles et al. (2010) suggested a biased random sampling algorithm which said that to achieve self organization, a dynamic approach of random sampling is used, so that the load is balanced in all available nodes. The load on the server will be shown in the virtual graph which concerns the task execution and completion within the network:

i. In a virtual graph, to indicate the work of the resources, it deletes an incoming edge, whenever a node will executes a job.

ii. When the work of the task is finished, the resources will be freed.

The accumulation and removal of tasks are done through "Random sampling algorithm". In the first step the node is selected randomly and then a neighbor node is selected, this process goes on but the last node is selected based on the load to be allocated. Other concept for the chosen of node for the allocation of load is based on the criterion like computing efficiency etc. One more concept, for selecting the node for allocation of load is that select the less loaded node that has the highest in degree. The efficiency of the load allocation is directly proportional to the walk length of b, i.e., if b increases the efficiency also increases. Here the "threshold value" of "b" is in general equal to "log n" practically. The job will be executed if the "current walk length" will be "equal and greater than" the "threshold value". Otherwise another neighbor node is selected randomly. But, after execution of the job, an edge which is incoming will be deleted. As the job is completed, an edge will be formed as of the "node" starting the process allotment to the "node" that was carrying out the job. At last, the directed graph is formed.

D. Zhang et al. (2009) suggested a"fast adaptive load balancing algorithm" which states that to split the "simulation region" into "sub-domains" "binary tree structure" is used. The load is balanced recursively between "sub domain" by constrict and extend "sub-domain" in the group. This technique can adjust the "sub-domains" with weighty loads and mould their load very quickly.

Dhinesh Babu L.D et al. (2013) anticipated an algorithm named "honeybee behavior inspired load balancing algorithm". In this algorithm, there are two kinds of bees "scout bees" and "forager bees" .The work of both is to find and reap the "food". First, the "scout bees" will find the "food" and when they find it they return to their place, does the "waggle dance" relay the

other bees' about the quantity and the quality of the food. After that, the "forager bees" go after the "scout bees" to the "food" and get the food to their native place. Following which, they again do the waggle dance to let the other bees know the quantity of food left. So, according to the concept of load balancing algorithm the honey bees are the tasks that are removed from the overloaded VMs. These removed tasks will be assigned to the under loaded VMs, according to priority basis, i.e. higher priority task will be assigned to the VM first. The advantage of this algorithm enhance the throughput, minimize the execution time but it disadvantage is that the lower priority task will remain in queue if there is explosion of high priority tasks.

Bin Dong et al. (2012) anticipated "a dynamic file migration load balancing algorithm". It is supported on "distributed architecture". But, there are certain problems faced by large file system that are migration of dynamic file or instead using distributed approach, algorithm follow centralized approach. So a new algorithm come into existence i.e., self acting load balancing (SALB) algorithm, which will avoid such problems. In the "parallel file system" "data management" plays an important role. In this, the data move between the "memory" and the "storage" devices. "Scalability", "availability" and network transmission etc. are some of the issues for balancing load in the parallel file system. SALB algorithm is as follows: Here the load in each I/O servers is different. There are some decision making algorithms that will help to choose the server for execution. One of the reasons to choose the dynamic and distributed approach as in case of centralized approach if the central server gets down whole system will be affected and the performance is degraded. Secondly, instead of central node, group decision maker is used which reduces the communication cost. But, in this "global optimization" is a main concern. Finally, the distributed approach is used, which helps in increase in availability and reliability.

Yunhua et al. (2010) proposed the "Heat Diffusion Based Dynamic Load Balancing": The author discussed that there are two heat diffusion algorithms; global and local. In a "distributed virtual environment", if the concurrent user tries to access load at the same time then that can cause problem. So, this problem can be resolved by "Heat diffusion based dynamic load balancing algorithm". In accordance to this algorithm, the "virtual environment" is created in which it is divided into many number of square "cells" and these "cells" are divided into number of objects. Within the cell, in every interaction, every node will send the load to its neighboring node and then the distinction among the "current nodes" to that of "neighboring node" will

transfer. So, this concept is concerned to “heat diffusion” process. The concept is as same as heat transference from “high to low” objects, whilst they are sited with each other.

Markus et al.(2010) anticipated the “Decentralized Scale-Free Network Construction and Load Balancing in Massive Multiuser Virtual Environments” which states that the “online environment” is the backbone for the interconnected machines which shows the concept of “overlay networks”. Also, according to him, for better technological developments, virtual online world is one such thing that makes the chance to the world. So the proposed network system developed hyper verse architecture that makes good viability and “balancing of load” to the “dynamic virtual environments”. This “Hyper verse architecture” is accountable for the appropriate hosting of the “virtual world”. Also, surface of the World is subdivided in to “small cells” by the “self organized load balancing method”, and this load balancing algorithm is handled by a “public server”. In this, cells different “hotspots” are considered so that the “absolute mass” of the object in the cell can be evaluated by the “public server”. When increasing the network load “Hotspot” accuracy is better. The algorithm proposed by (Markus et al., 2010) finds the “number of links” that is allocated to every “node” while combining the “network” as well as cannot avoid the overloaded nodes. The benefits are: the “network” becomes “reliable”, “resilience”, “efficient routing”, and “fault tolerant”. The consequences of the proposed algorithm is the “overload ratio” at the starting is more due to that “public servers” are primarily positioned “randomly” so only a little time is used for balancing the load.

Kousik et al. (2012) anticipated **“**Genetic Algorithm Based Load Balancing” algorithm. Its main aim is to “balance the load” in cloud environment by minimizing the makespan. While trying to lessen the “makespan” of a job the genetic algorithm flourish to “balance the load” of the “cloud infrastructure”. According to the network load “Genetic” based concept follow some “rules” & “randomization” effectively.

Anamika et al. (2013) proposed this meta-heuristic ant colony optimization algorithm for load balancing. For optimal solutions several number of tasks and systems are used in it the heuristic algorithm. In order to process the job, the concept “ACO” is based on character of actual “ants” that will form the set of connections known as network. It is proposed by (Dorigo, 1991).For finding foods from source to nest the ants are moving in a path. The liquid that left for other ants to follows is known as “pheromone”. Another “ants” will go after the “pheromones” to reach the food sources; also the amount of pheromone will decide the superiority and amount of food

available. So, in the ant will move in such a way that they form a graph where all the nodes are connected. Also, if the optimal solutions are not found then they move randomly. The disadvantage of this algorithm, is the network overhead due to large number of ants, confusion in number of ants ,and last is only availability of node is being considered , and no other factors are considered.

3. Description of Broad Area

"Load Balancing" is the concept of dispensing of the task to the set of computers in a group of systems (network). It will help to improve the "Performance" of the cloud environment via reducing response time, makespan time and increasing throughput. This technique avoids too much overload and reduces the" total waiting time" of the resources. Also, in this, among servers traffic is divided so that without maximum delay data can be sent and received. The division of workload dynamically to the servers is an important issue in the cloud environment. Here, workload means the task or the job and the workload of the machine means how much time is required to "execute" all the jobs which is allocated to the "machine". So, "Load Balancing" is defined as the procedure of shifting of workload among the processors by enhancing the performance of the system. For achieving maximum client satisfaction, by distribution of the workload for higher resource utilization ratio is an advantage of "load balancing". Through "load balancing" the equivalent quantity of work is done by every virtual machine in cloud system which will maximizes the "throughput" and minimizes the "response time". So, the balancing the load is a crucial feature to improve the "performance" of "cloud service provider". Here, in Figure 7, overview of the balancing of load in Cloud Computing Environment is shown.

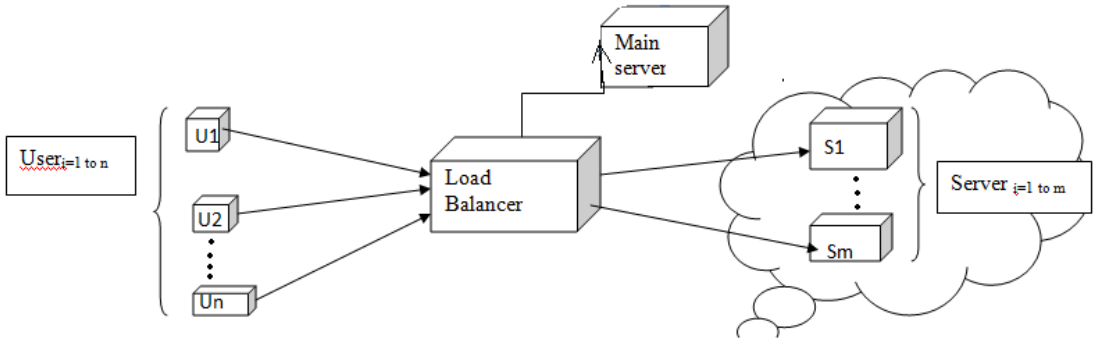

Figure 7: Balancing Load in "Cloud Computing" Environment

3.1 Need for Load Balancing in Cloud Environment

The main aspiration is to attain "high user satisfaction" and "resource utilization ratio", i.e. done by the load balancer which dispense the dynamic work load amongst different systems or nodes in the cloud environment so that a single system or node is not overburdened this will enhance the "overall performance" of the node or system. However the fact is that the "load balancing"

must be done in a way that will help to attain optimal resource utilization and reduce the resource consumption. Figure 8 depicts the role of load balancer. Some other benefits of balancing the load are enabling scalability, applying fail-over, over-provisioning, avoiding bottlenecks, achieving Green Computing and reducing response time in clouds. There is some feature accountable for it that are:

i. **Energy Consumption is reduced**: By avoiding overheating of nodes or virtual machines, Load balancing helps in reducing the amount of energy consumption due to desirable workload.

ii. **Carbon Emission is limited**: Carbon is emitted due to energy consumption as they both are directly proportional to each other. So, as the energy consume is less it will lead to green computing means reduction in carbon emission.

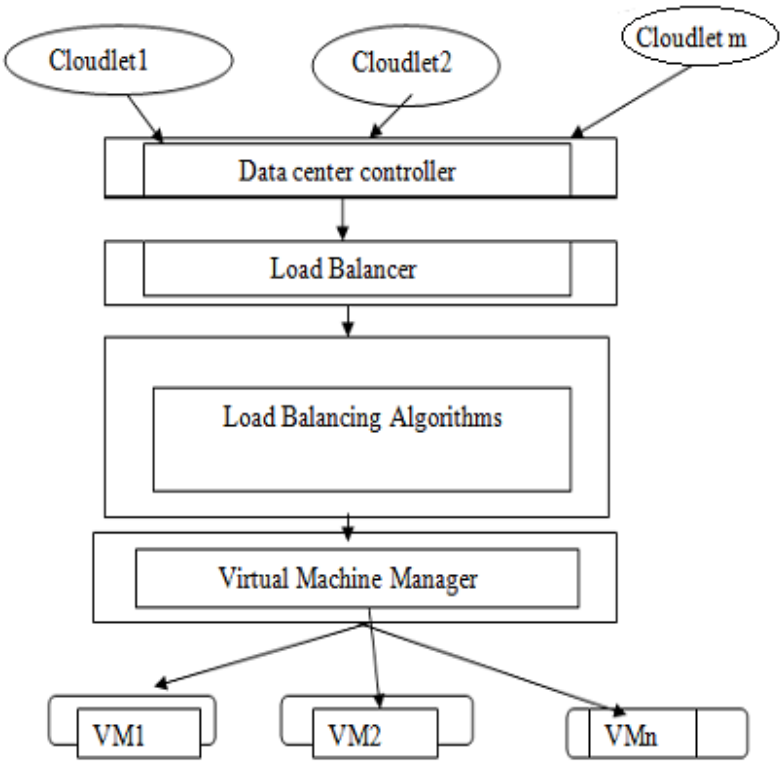

Figure 8: Load Balancer

3.2 Categories of balancing the load on the basis of cloud environment are:

Following Figure 9 shows the categories of Load balancing:

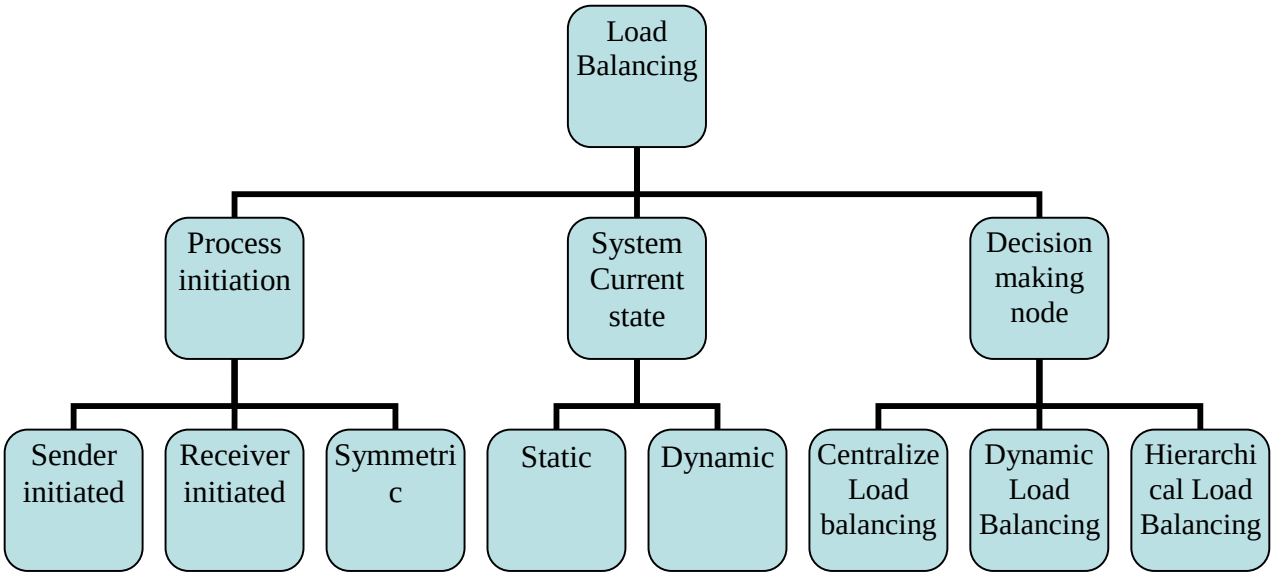

Figure 9: Categories of balancing the load (Nitin et al. (2015))

Three categories of load balancing based on process initiation can be categorized into:

1. **Sender Initiated**: In the "Sender Initiated Algorithm", sender will start the process. In this client sent the request and the receiver will receive the workload sent by the client i.e. the source or the sender has started the procedure.

2. **Receiver Initiated**: In the "Receiver Initiated Algorithm", the target or the receiver will start the process. In this, the receiver will send the acknowledgement request to the "source or sender" that is already prepared to allocate the "workload".

3. **Symmetric:** It combines both the "sender initiated" & "receiver initiated" categories of "load balancing algorithm".

Two categories of "balancing the load, based on the current state of the system":

1. Static Load balancing: In Static load balancing algorithms, there is no need of any prior information regarding present state as prior information about the memory; storage etc is already there with the load balancers. There are certain consequences in this load balancing i.e., a sudden failure of the system or also shifting of task can't be possible. One of the good examples of static load balancing is round robin algorithm and its improved version is weighted round robin algorithm.

2. Dynamic Load Balancing: In Dynamic load balancing algorithms, earlier existing information about the node is not required whereas the decision is taken on the current state of the node. A dynamic approach in the load balancing algorithm will enhance the performance and

also make tolerance to some faults. Also, some policies are there that are utilize in "dynamic load balancing algorithm" which are as follows:

1. **Transfer Policy:** Transfer policy is the transferring the job from local node to remote node in a dynamic manner.

2. **Selection Policy**: This policy will specify the selection of the processors for exchange of

3**. Location Policy**: This policy or strategy states the selection of a destination node for transferring of task.

4. **Information Policy**: For dynamic approach knowledge need to be collected about the node in the system for balancing the load. There are certain information policies that are:

i. **Load Estimation Policy:** Estimation of the total workload on the processor is predicted by this "policy".

ii. **Process Transfer Policy:** This policy will suggest whether the task will be executed locally or remotely.

iii. **Priority Assignment Policy:** According to this policy, priority of the task is assigned to execute either locally and remotely.

iv. **Migration Limiting Policy:** This policy will set the limit that the largest number of time that the task or job can migrate from one system to another.

Figure 10 shows the interaction among these policies in dynamic load balancing.

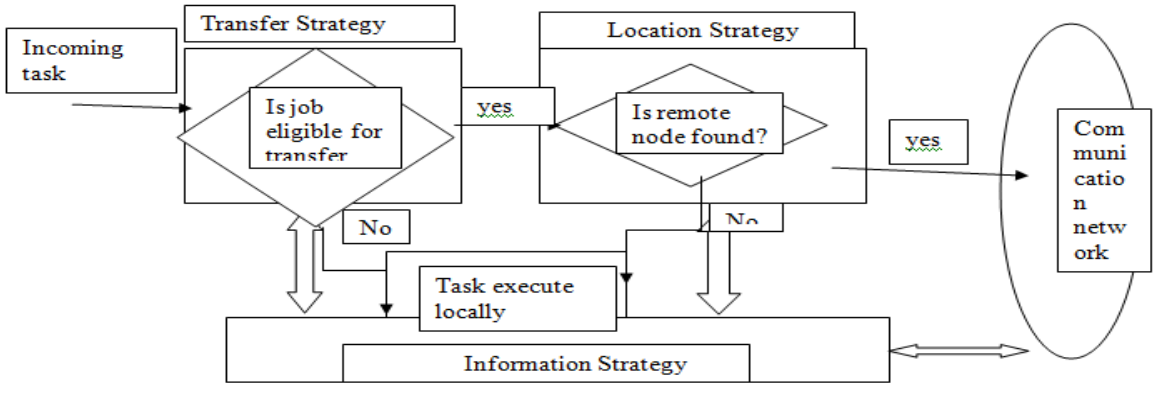

Figure 10: In dynamic load balancing algorithm interaction among these policies.

There are three categories of "load balancing algorithms" that depends on the "decision making":

1. Centralized Load balancing: In this, a "single node" will allocate and schedule the job. This centralized node will consist of all the information of the network of cloud that can be functional to both static and dynamic approach for harmonizing the load. Through this type of load balancing the time required to evaluate or analyze the different resources is reduced but has the disadvantage that it creates a lot of load on the centralized, as well as do not afford "fault tolerance" .Through this technique revitalization is also problematic in case of collapse.

2. **Distributed Load Balancing**: In this type of load balancing, there is no centralized node for decisions making instead multiple nodes are used for load balancing due to which this load balancing algorithm is fault tolerant. So, for this every node must have knowledge of each other to get better performance. .

3**. Hierarchical Load Balancing**: In "hierarchical load balancing algorithm", different stages of the cloud are implicated. In "hierarchical load balancing", slave mode operation technique is used that can be employed using "tree data structure", where "parent-child relationship" is shown by the tree structure. For getting statistics of slave nodes, master uses light weight agent process and "parent node" obtain the conclusion based on the collection of information.

3.3 Qualitative metrics of Load Balancing

In order to find the best algorithm "Qualitative metrics" consists some parameters which are useful among them. There are various "qualitative metrics" and "parameters" that are well thought-out for balancing load in "cloud computing" is as follows:

i. **Throughput:** Throughput is the "total number of task" executed in a given "amount of time". For better "performance" of the machine, "throughput" must be high.

ii. **Associated Overhead:** For executing "load balancing algorithm "the amount of overhead produced should be minimum.

iii. **Fault tolerant:** "The load balancing algorithm" should be able to execute well even in the case of failures.

iv. **Migration time:** "Migration time" is the duration by the job to migrate from one machine to other and it must be lowest for improved "performance" of the machine.

v. **Response time**: Response time is the time to execute the specific task in a minimum amount of time.

vi. **Resource Utilization**: "Utilization" of the resources fully improves the "performance' of the machine. How much resources are being utilized is termed as resource utilization.

vii. **Scalability**: Scalability means to scale up and down the resources as the need of the "load balancing algorithm".

viii. **Performance:** The main factor is to increase in performance of the system. This is only possible if an only if all the above parameters are satisfied.

3.4 Load Balancing Challenges in Cloud Computing

There is still some confronts remain unsettled by the "scientific community", though "cloud computing" has been broadly accepted but research are still in there early stages.

Some "challenges in load balancing" are as follows:

1. **Automated Service provisioning**: The main challenge in load balancing is the automation in which of the resources which can be allocated or released mechanically according to the need.

2. **Virtual machine Migration**: "Virtual machine migration" is the challenge in load balancing as it means to migrate the task to the under loaded machine from the overloaded machine. While migration you have to take certain parameters into consideration and also there are various types of virtual migration. To allocate the task in a "datacenter" or "set of datacenters" is the main objective of virtual machine migration.

3. **Energy Management**: Energy saving and energy management is one of the crucial challenge for load balancing algorithms. To achieve green computing in cloud environment energy saving is extremely needed. There is all time need for "energy efficient algorithm" that reduces the "resource consumption" but maintain fine "performance".

4. **Stored Data Management**: There is an exponential growth in the last few years due to that large amount of data needs to be stored. This data may belong to any corporation or any person. A bigger issue for cloud computing is the management of data storage for corporation or person. So, the question is that how to "distribute" the data to the storage for better "storage of data" with the "fast access"?

5. **Emergence of Small Data Centers for Cloud Computing**: Due to more advantageous, inexpensive and less energy consumer "small datacenters" have some profit over "large data centers". Leading to geo-diversity computing "small providers" can convey "cloud computing services". To make certain sufficient "response time" with an optimal sharing of resources, balancing the load will become a trouble on a worldwide. So, for improving the overall "performance" of machine, "load balancing" is used to distribute the "higher processing load" to

“lower processing nodes”. To disseminate the load evenly in the midst of the nodes balancing of load is required in cloud computing environment.

• To attain a “high satisfaction of user” & “proper utilization of resources” balancing of load helps in reasonable allotment of computing resource. “High resource utilization” & “Proper load balancing” helps in reducing “resource consumption” and in “implementing fail over”, “scalability”, & ” avoiding bottlenecks”.

• “Load balancing” is a concept that split the traffic among all servers, so the tasks can be executed with no delay with “load balancing”. Also, “load balancing” helped networks and resources by providing” minimum response time” and “Maximum throughput”.

• There are many algorithms in cloud environment that will traffic proper loads amongst the servers. There are two groups in cloud computing environment for load balancing algorithms: One is “Batch mode heuristic scheduling algorithms (BMHA)” and the other is “Online Mode Heuristic algorithms (OMHA)”. In “BMHA scheduling algorithm”, first step is to combine all the jobs, whenever they will reach in the system and then will start after a rigid time period.

• Some of the examples of “BMHA based algorithms” are: ”First come first serve basis(FCFS)”, “Round Robin scheduling algorithm (RR)”, “Min Min algorithm” & “Max Min algorithm”. In “On-line mode heuristic scheduling algorithm”, when all the tasks are arriving in the system, these Jobs are scheduled. The “Online Mode Heuristic Scheduling Algorithms” are more suitable and healthier for a cloud environment.

• While developing a “load balancing algorithm” following points need to be taken care of that are: estimation of the proper load, comparison among load, performance of the considerate machine, stability of all the systems, nature of work to be transferred etc. The necessary thing is the selection of the nodes, also to calculate the load of machine “CPU load”, “amount of memory” required combine together.

• Daily life example for “load balancing” is the “websites” like in case of absence of load balancing users may experience certain problems like delays , timeouts etc.

3.5 Virtualization

Virtualization is the environment that is not “real” although it offers the whole thing “real”. It is the “software” implemented as the system that will act exactly like the actual system. It is very important part of cloud environment as the user of the virtual system can use different “applications”,” infrastructure” or “services” of the cloud.

Different kinds of virtualization in the cloud environment:

1) “Full virtualization”

2) “Para virtualization”

1) **Full Virtualization**: In this “virtualization”, only a section of hardware and is emulated and rest of the hardware is used for real host system. Due to which as you bypass the emulation layer the performance increases

2) **Para virtualization**: In Para virtualization, hyper call API is used to communicate with the hypervisor for instruction dispatch and therefore, there is no hardware simulation/emulation done by the virtual machine. Example: Xen –Para Virtualization

4. Problem Identification and Objective of Study

Cloud Computing is a network system based on internet in which various scalable resources or services are provided by the "cloud provider" to the "cloud users" over the internet. Since, cloud computing environment is rapidly growing in number with kind of resources and services being provided, also we know that the resource are connected over larger geographical area that have to be utilized efficiently, there is a requirement for proficient balancing of load and scheduling of task. Load Balancing is an NP Hard problem, which can be solved using meta-heuristic nature inspired load balancing algorithms. So, the objective of my research work is to enhance or improvise the performance of the cloud environment by optimizing some load balancing technique. Load Balancing has two key features:

1) To schedule the task and
2) To manage the resource utilization.

So, to enhance or improvise the performance of cloud environment there is a need to optimize either task scheduling or manage the resource utilization.

Objectives of Study

1) To enhance the performance of the cloud environment by optimizing some nature inspired load balancing algorithm.
2) To reduce the execution time, makespan and cost of the algorithm used for load balancing.
3) To fetch all the available virtual machines in the datacenter.
4) To balance the load equally among all virtual machines.
5) To retrieve the processing capacity of the available virtual machines.

5. Methodology to be adopted

Following are the steps or phases to achieve the objective of the research:

Step 1: Study the concept of load balancing which states its definition, scenario in the cloud environment.

Step 2: Survey on the literature of various load balancing techniques.

Step 3: Identify the areas and qualitative metrics on which the performance of the system in the cloud environment is based.

Step4: Draw the flowchart according to the scenario identified or proposed.

Step 5: Identify the tools or software to implement the proposed work.

Step 6: Compute the performance of the proposed work using graphs and charts.

Step 7: Compare the result with the earlier work which states the better performance of the proposed work.

6. Proposed/expected outcome of the research

The research work on balancing the load in the environment of cloud involves the enhancement of the "performance" in terms of the below stated metrics:

The "performance" of the "load balancing algorithm" is based on "Execution time", "Makespan", "Cost" and "Degree of Imbalance".

1) Execution Time: Execution time is defined as in how much time the task get executed by virtual machine.

Execution time= (Length of task /Processor Capacity)+(Size of data file/Bandwidth ... (1)

2) Cost: Here cost is actually the processing cost, which states the amount or cost for which the user has utilized the set of resources for certain period of time. The cloud providers' main motive is to get profit or add on revenue as much as possible whereas the cloud user's main motive is to get the resources in fewer amounts for more period of time. Equation 1 represents the cost of the executing task.

Cost= sum of all task's (Cost * Execution time) … (2)

3) MakeSpan: "MakeSpan" is defined as the "execution time" of the last task, when all the tasks are scheduled. The motive is to minimize the makespan of specific cloudlet.

MakeSpan= $\max_{task}$(Finishing time) … (3)

4) Throughput: "Throughput" is defined as no. of task executed divided by total execution time.

Throughput= (no. of tasks executed)/(total execution time) **…..(4)**

7. Proposed Time Frame (Gantt chart)

Activities	SEP-DEC (2014)	JAN-JUNE (2015)	JULY-DEC (2015)	JAN-JUNE (2016)	JULY-DEC (2016)	JAN-JUNE (2017)	JULY-DEC (2017)	JAN-JUNE (2018)	JULY-DEC (2018)	JAN-JUNE (2019)
Study the concept of load balancing										
Survey the literature										
Identify the areas and metrics for performance										
Evaluate the performance and simulate using tools and software										
Compare the results with the earlier one to state that the performance of my work is better than the earlier one										

References

A.Bhadani and S. Chaudhary(2010), "Performance evaluation of web servers using central load balancing policy over virtual machine on cloud*"*, proceedings of third Annual ACM.

Amit Kumar, Harish Chandra Maurya, Rahul Misra,(2013), *"*A Research Paper on Hybrid Intrusion Detection System", International Journal of Engineering and Advanced Technology (IJEAT),ISSN: 2249 – 8958, Volume-2, Issue-4.

Anamika Jain, Ravinder Singh (2013), "Review of Peer to Peer Grid Load Balancing Model Based on Ant Colony Optimization with Resource Management*"*, IJARCSSE, Volume 3, Issue4.

Aurobindo Sundaram (1996),"An Introduction to Intrusion Detection", Volume2, Issue 4, pp-3-7.

Bala, Anju and Chana, Inderveer (2011), "A survey of various workflow scheduling algorithms in cloud environment", 2nd National Conference on Information and Communication Technology (NCICT), pp-26-30.

Bin Dong, Xiuqiao Li, Qimeng Wu, Limin Xiao, Li Ruan(2012), "A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers*",* J. Parallel Distribution Computing,Volume 72, pp-1254–1268.

B. Grobauer, T. Walloschek, and E. Stöcker (2011), "Understanding Cloud Computing Vulnerabilities," 2011 IEEE Security and Privacy, pp. 50-57.

Chaczko, Zenon and Mahadevan, Venkatesh and Aslanzadeh, Shahrzad and Mcdermid, Christopher (2011), "Availability and load balancing in cloud computing", International Conference on Computer and Software Modeling, Singapore, Volume 14.

Chaudhari, Anand and Kapadia, Anushka (2013), "Load Balancing Algorithm for Azure Virtualization with Specialized VM",algorithms, Volume 1,pages 2.

Dhinesh Babu L.D, P. VenkataKrishna (2013*)*, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", Applied Soft Computing, Volume 13 , pp-2292–2303.

Dongliang Zhang, Changjun Jiang, Shu Li (2009), "A fast adaptive load balancing method for parallel particle-based simulations", Simulation Modelling Practice and Theory, Volume 17, pp-1032–1042.

Hitesh Bheda, Hiren Bhatt(2015)," An Overview of Load balancing Techniques in Cloud Computing Environments*"*, International journal of Engineering and Computer Science Volume 4, pp.9874- 9881.

J. M. Galloway, K. L. Smith, and S. S. Vrbsky (2011), “Power aware load balancing for cloud computing,” in Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp.19–21.

Kousik Dasgupta, Brototi Mandal, Paramartha Dutta (2012), “Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach”, Elsevier (C3IT) .

Markus Esch, Eric Tobias(2010), “Decentralized scale-free network construction and load balancing in Massive Multiuser Virtual Environments”, in: Collaborative Computing: Networking, Applications and Work sharing, Collaborate Com, 6th International Conference on, IEEE, pp. 1–10.

Mohammad Sajid, Zahid Raza,(2013) “Cloud Computing: Issues & Challenges, International Conference on Cloud, Big Data and Trust”, Nov ,pp-13-15, RGPV.

M. Randles, D. Lamb, and A. Taleb-Bendiab(2010), "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", Proceedings IEEE International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, pp.551-556.

Nayandeep Sran, Navdeep Kaur(2013) , “Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing ” ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726, Volume 2 Issue 1 ,January, pp-.60-63

Nidhi Jain Kansal, Inderveer Chana(2012), "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, ISSN (Online): 1694-0814 , Volume 9, Issue 1,pp-238-246.

Nitika, Shaveta, Gaurav Raj(2012); “Comparative Analysis of Load Balancing Algorithms in Cloud Computing”, International Journal of Advanced Research in Computer Engineering & Technology, ISSN: 2278 – 1323, Volume 1, Issue 3,pp-120-124.

Nitin Kumar Mishra ,Nishchol Mishra(2015),”Load Balancing Techniques :Need ,Objective and Major Challenges in Cloud Computing –A Systematic Review” International Journal of Computer Applications, Volume 131,Issue 18,pp-11-19.

Nusrat Pasha, Dr. Amit Agarwal and Dr.Ravi Rastogi(2014),“Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment” International Journal of Advanced Research in Computer Science and Software Engineering , ISSN: 2277 128X,Volume 4, Issue 5,pp-34-39.

Poonam Devi, Mr. Trilok Gab (2013) ”Implementation of cloud computing by using short job scheduling” International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3,Issue 7,pp-178-183.

Santosh Kumar and R. H. Goudar (2012),”Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey” International Journal of Future Computer and Communication, ISSN: 2010-3751 ,Volume 1, No. 4,pp-356-360.

Shahrzad Aslanzadeh, Venkatesh Mahadevan, Christopher Mcdermid (2011), “Availability and Load Balancing in Cloud Computing” International Conference on Computer and Software Modeling IPCSIT volume 14 IACSIT Press, Singapore , pp-134-140.

S. Sethi, A. Sahu, and S. K. Jena (2012), “Efficient load balancing in cloud computing using fuzzy logic,” IOSR Journal of Engineering, Volume 2, no. 7, pp.65–71.

S. Wang, K. Van, W. Liao, and S. Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", Proceedings of the 3rd IEEE International Conference on Computer Science and Information , Chengdu, China, September 2010, pp. 108-113.

T. Kokilavani and Dr. D.I. George Amalarethinam(2011), “Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing” International Journal of Computer Applications ,Volume 20,No.2, pp.0975-8887.

Yunhua Deng, Rynson W.H. Lau(2010), “Heat diffusion based dynamic load balancing for distributed virtual environments”, in: Proceedings of the17th ACM Symposium on Virtual Reality Software and Technology, ACM, pp. 203–210.

Z. Nine, M. SQ, M. Azad, A. Kalam, S. Abdullah and R. M. Rahman (2013), “Fuzzy logic based dynamic load balancing in virtualized data centers” In fuzzy system (FUZZ), IEEE International conference on, pp. 1-7.

URL:
http://www .loadbalancing.org/.